Is_Cereal_Killer

Sophia Waxman

2025-11-04

```r
#How many cereals does each manufacturer make?
table(data$mfr)
```

```
##
##  A  G  K  N  P  Q  R
##  1 22 23  6  9  8  8
```

```r
#How many cereals are hot vs. cold?
table(data$type)
```

```
##
##  C  H
## 74  3
```

```r
#How heavy are the cereals on average on each shelf?
shelf1 <- subset(data, shelf == "1")
shelf2 <- subset(data, shelf == "2")
shelf3 <- subset(data, shelf == "3")

mean(shelf1$weight/shelf1$cups)
```

```
## [1] 1.159961
```

```r
mean(shelf2$weight/shelf2$cups)
```

```
## [1] 1.159131
```

```r
mean(shelf3$weight/shelf3$cups)
```

```
## [1] 1.683875
```

```r
#Which cereal has the highest amount of sugar per weight (highest sugar density)?
data <- mutate(data, sugarden = data$sugars/data$weight)
data$name[which.max(data$sugarden)]
```

```
## [1] "Golden Crisp"
```

```r
#Which cereal manufacturer produces the most fibrous cereal on average?
A <- subset(data, mfr == "A")
G <- subset(data, mfr == "G")
K <- subset(data, mfr == "K")
N <- subset(data, mfr == "N")
P <- subset(data, mfr == "P")
Q <- subset(data, mfr == "Q")
R <- subset(data, mfr == "R")

DF <- data.frame(meanfiber = c(mean(A$fiber), mean(G$fiber), mean(K$fiber), mean(N$fiber),
                   mean(P$fiber), mean(Q$fiber),mean(R$fiber)),
                 row.names = c("A", "G", "K", "N", "P", "Q", "R"))

rownames(DF)[which.max(DF$meanfiber)]
```

```
## [1] "N"
```

```r
# Which cereals are the least healthy?
data2 <- data %>%
  mutate(
    fiber_z = scale(fiber),
    protein_z = scale(protein),
    vitamins_z = scale(vitamins),
    sugars_z = scale(sugars),
    calories_z = scale(calories),

    health_score = ((fiber_z + protein_z + vitamins_z) - (sugars_z + calories_z))/cups
  )

arranged <- data2 %>%
  arrange(desc(health_score)) %>%
  select(name, mfr, cups, fiber, protein, vitamins, sugars, calories, health_score)

head(arranged, 5)
```

```
## # A tibble: 5 x 9
##   name       mfr    cups fiber protein vitamins sugars calories health_score[,1]
##   <chr>      <chr> <dbl> <dbl>   <dbl>    <dbl>  <dbl>    <dbl>            <dbl>
## 1 All-Bran ~ K      0.5     14       4       25      0       50            21.3
## 2 100% Bran  N      0.33    10       4       25      6       70            19.9
## 3 All-Bran   K      0.33     9       4       25      5       70            19.3
## 4 Grape-Nuts P      0.25     3       3       25      3      110             5.39
## 5 Total Who~ G      1        3       3      100      3      100             5.22
```

```r
tail(arranged, 5)
```

```
## # A tibble: 5 x 9
##   name       mfr    cups fiber protein vitamins sugars calories health_score[,1]
##   <chr>      <chr> <dbl> <dbl>   <dbl>    <dbl>  <dbl>    <dbl>            <dbl>
## 1 Frosted F~ K      0.75     1       1       25     11      110            -4.16
## 2 Cinnamon ~ G      0.75     0       1       25      9      120            -4.80
## 3 Fruity Pe~ P      0.75     0       1       25     12      110            -5.02
```

```
## 4 Mueslix C~ K      0.67     3       3       25      13      160             -5.18
## 5 Cap'n'Cru~ Q      0.75     0       1       25      12      120             -5.70
```

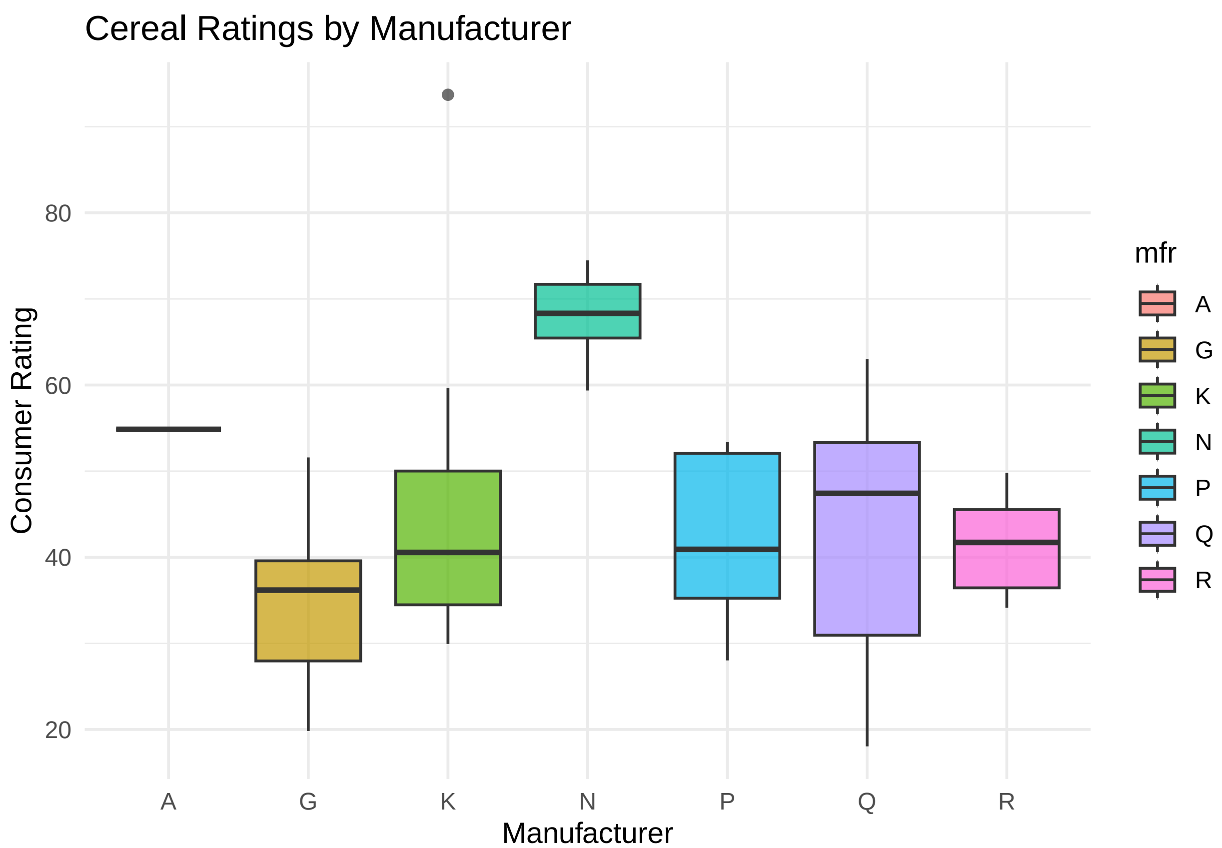
```r
# Does the rating of the cereal differ by manufacturer?
m2 <- lm(rating ~ mfr, data=data)

anova2 <- aov(m2)
summary(anova2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## mfr          6   5524   920.7   6.804 1.03e-05 ***
## Residuals   70   9473   135.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(anova2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = m2)
## 
## $mfr
##           diff         lwr       upr     p adj
## G-A -20.365065 -56.4729285 15.742798 0.6103726
## K-A -10.812455 -46.8861733 25.261264 0.9698076
## N-A  13.117650 -25.0260311 51.261331 0.9417294
## P-A -13.145173 -50.3695943 24.079248 0.9341922
## Q-A -11.934927 -49.3912786 25.521424 0.9593390
## R-A -13.307920 -50.7642713 24.148432 0.9323420
## K-G   9.552611  -0.9786463 20.083868 0.1003195
## N-G  33.482715  17.2181951 49.747236 0.0000006
## P-G   7.219892  -6.7533586 21.193143 0.7023999
## Q-G   8.430138  -6.1497268 23.010003 0.5820387
## R-G   7.057145  -7.5227196 21.637010 0.7616552
## N-K  23.930105   7.7415282 40.118681 0.0005307
## P-K  -2.332718 -16.2174990 11.552063 0.9986352
## Q-K  -1.122472 -15.6175701 13.372625 0.9999849
## R-K  -2.495465 -16.9905629 11.999632 0.9984337
## P-N -26.262823 -44.8750337 -7.650612 0.0010854
## Q-N -25.052577 -44.1244179 -5.980737 0.0029503
## R-N -26.425570 -45.4974107 -7.353729 0.0014161
## Q-P   1.210246 -15.9493646 18.369856 0.9999913
## R-P  -0.162747 -17.3223574 16.996863 1.0000000
## R-Q  -1.372993 -19.0300862 16.284101 0.9999846
```

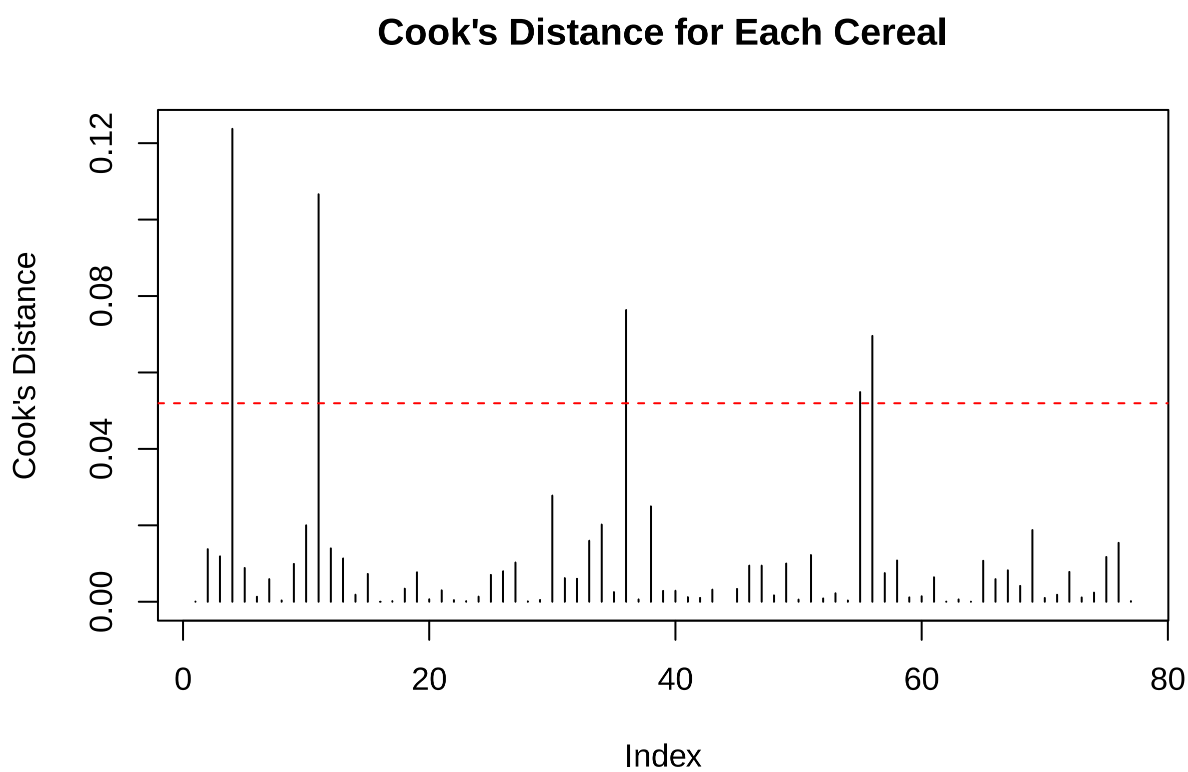
```r
ggplot(data, aes(x = mfr, y = rating, fill = mfr)) +
  geom_boxplot(alpha = 0.7) +
  theme_minimal() +
  labs(title = "Cereal Ratings by Manufacturer",
       x = "Manufacturer",
       y = "Consumer Rating")
```




```r
#Outliers
# Cook's distance
cooks <- cooks.distance(anova2)

# Find the highest value
outlier_index = which.max(cooks)

# View that observation
data[outlier_index, c("name", "mfr", "rating", "sugars", "calories")]
```

```
## # A tibble: 1 x 5
##   name                      mfr   rating sugars calories
##   <chr>                     <chr>  <dbl>  <dbl>    <dbl>
## 1 All-Bran with Extra Fiber K       93.7      0       50
```

```r
# Optional plot
plot(cooks, type = "h",
     main = "Cook's Distance for Each Cereal",
     ylab = "Cook's Distance")
abline(h = 4 / length(cooks), col = "red", lty = 2)
```




```r
# Remove the outlier
data_no_outlier <- data[-outlier_index, ]

# Re-run ANOVA
anova_no_outlier <- aov(rating ~ mfr, data = data_no_outlier)
summary(anova_no_outlier)
```

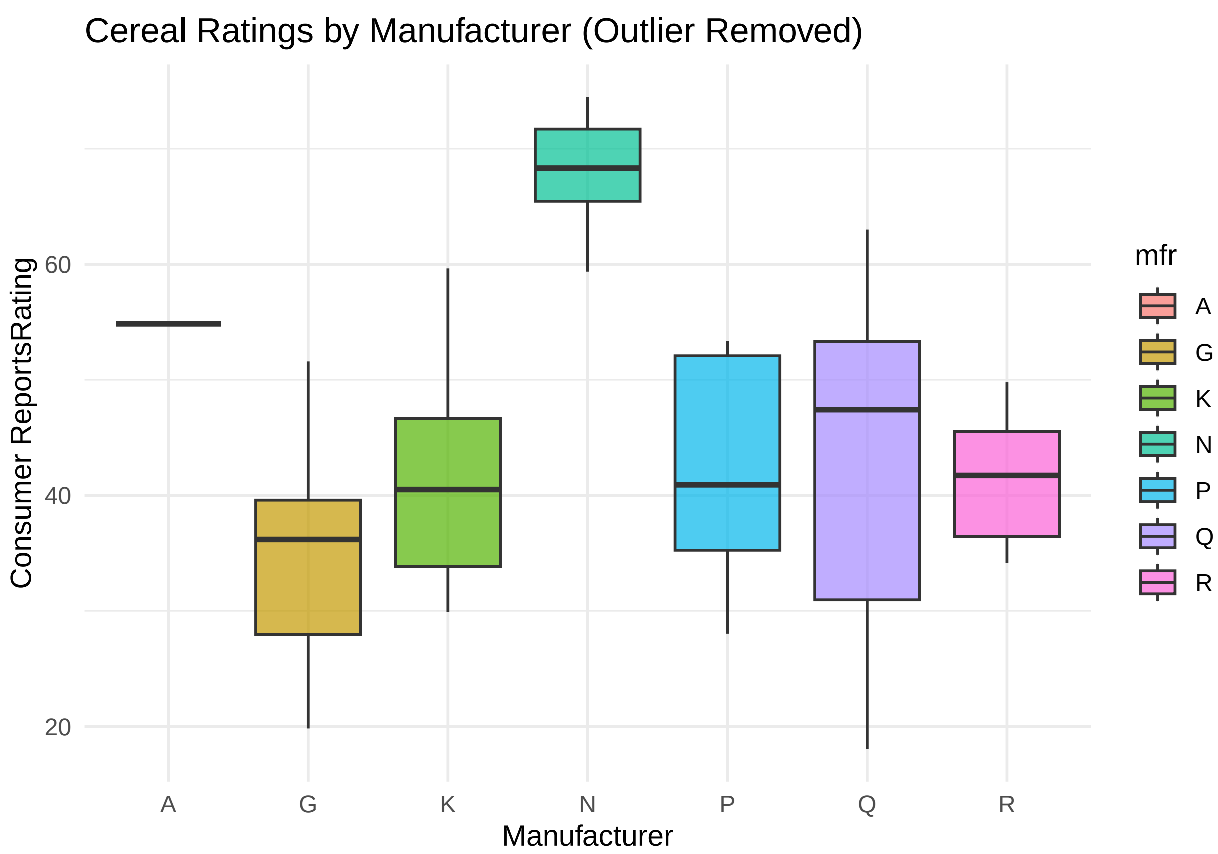
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## mfr          6   5464   910.6   9.114 2.45e-07 ***
## Residuals   69   6894    99.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Compare with original model
summary(anova2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## mfr          6   5524   920.7   6.804 1.03e-05 ***
## Residuals   70   9473   135.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ggplot(data_no_outlier, aes(x = mfr, y = rating, fill = mfr)) +
  geom_boxplot(alpha = 0.7) +
```

```r
  theme_minimal() +
  labs(
    title = "Cereal Ratings by Manufacturer (Outlier Removed)",
    x = "Manufacturer",
    y = "Consumer ReportsRating"
  )
```




```r
#Are sugar and fiber inversely correlated? (regression model)
suga <- data$sugars != "-1"
sugar <-data$sugars[suga]
fiber <- data$fiber[suga]
cor.test(sugar, fiber)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sugar and fiber
## t = -1.2053, df = 74, p-value = 0.2319
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.35316657  0.08949582
## sample estimates:
##        cor
## -0.1387595
```

```r
#Is the type of cereal correlated with potassium content? (regression model)
pot <- data$potass != "-1"
potato <-data$potass[pot]
hotcold <- data$HotCold[pot]
cor.test(data$HotCold, data$potass)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$HotCold and data$potass
## t = 0.69352, df = 75, p-value = 0.4901
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1467781  0.2984675
## sample estimates:
##        cor
## 0.07982503
```